Indoor-Outdoor Image Classification using Mid-Level Cues

Yang Liu* and Xueqing Li†
*Department of Computer Science and Technology, Shandong University, Jinan, Shandong, P.R.C.
E-mail: lawyoung@sdu.edu.cn Tel: +31647260761
†Department of Computer Science and Technology, Shandong University, Jinan, Shandong, P.R.C.
E-mail: xqli@sdu.edu.cn Tel: +8613806412620

***Abstract*—Classifying an image into indoor/outdoor image category is very difficult due to vast range of variations in both of these scene categories. Most previous indoor-outdoor classification approaches utilize the simple statistics of the low-level features, such as colors, edges and textures. In this paper, we incorporate mid-level information to obtain superior scene description. We hypothesize that pixel based low-level descriptions are useful but can be improved with the introduction of mid-level region information. Experiments show that, while using mid-level features, it produces comparable result with that using low-level features. When combined with low-level features, the classification result get improved.**

I. Introduction

Semantic scene classification is defined as the process of automatically categorizing images into a discrete set of semantic classes such as indoor and outdoor. Scene classification is a fundamental problem in image understanding,which have a high potential for improving the performance of other computer vision applications. In [1], different color constancy methods are selected for indoor and outdoor images each,which highly improves the final results.

As humans, we are extremely proficient at perceiving natural scenes and understanding their contents. However, the automatic classification of digital photographs into indoor/outdoor image categories is not an easy task for computers owing to their variability, ambiguity, and the wide range of illumination and scale. Ideally, a classification system should provide accurate performance while ignoring these variations. To bridge this “ semantic gap”, a lot of methods have been proposed. [1-3] utilize the simple statistics of the low-level features, such as colors, edges and textures. Oliva and Torralba [4] represent the whole scene with a very low dimensional termed Spatial Envelope and classify the images using nearest neighboring. All these approaches consider the scene as an individual object and investigate the “ semantic gap” between the low-level visual features and the high-level concept classification.

In this paper, we are going to explore the feature description by using spatial information from mid-level cues. Therefore, we introduce a superpixel-based region descriptor and apply it to indoor/outdoor image classification. The superpixels of the images, as shown in Fig. 1, are usually the semantic objects or part of them, which make them an ideal candidate for image classification. The objects usually exist in certain scene and can provide meaningful information for image indoor/outdoor classification. For example, for images representing outdoor scenes, sky, trees and buildings usually appear in them. This rationale holds also for other concepts where tables, chairs and walls can be found. Pixel based descriptors are widely used in image classification tasks due to their performance for image description. However, the use of mid-level descriptors is also important for a better scene characterization. The aim of this paper is to extend the low level descriptors towards mid-level region descriptors to improve the image classification results.

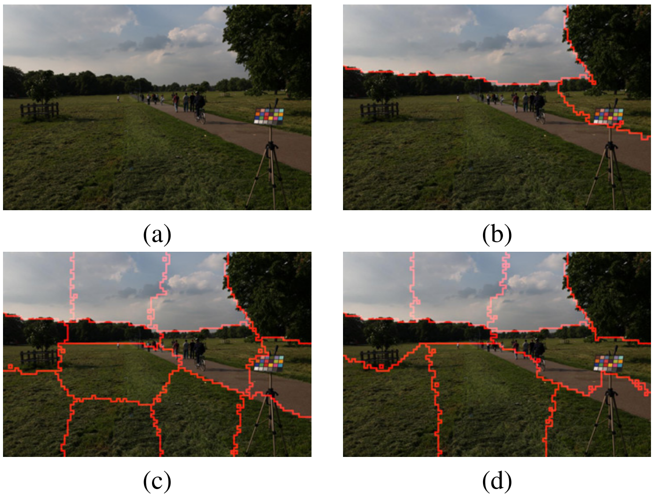

Fig. 1. Describing an image with superpixels. (a) is the original image; (b-d) show the superpixels generated from Normalize Cut method by varying different parameters.

The remainder of this paper is organized as follows. In Section II, we give a brief overview of normal image indoor/outdoor classification method. In Section III, we describe how to generate mid-level cues from superpixels in detail. In Section IV, experimental results are given followed by our conclusion in Section V.

II. Indoor/Outdoor Classification

In this section, we will first give the normal pipeline for indoor and outdoor classification, and then talk about the low-level features and the classifier that we incorporate into this pipeline. The pipeline for indoor/outdoor classification is shown in Fig. 2 and consists of two major steps: (1) feature extracting; (2) scene classifying. For each image, a set of low-

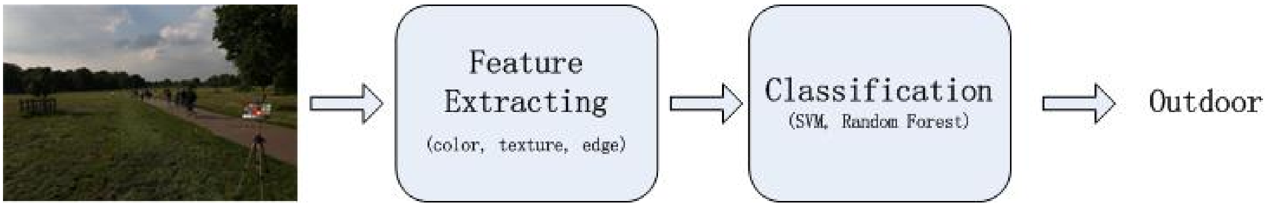

Fig. 2. A normal pipeline for image indoor/outdoor classification.

level features, related to color, texture, and edge distribution is extracted. Then, the extracted features, organized in a feature vector, are fed into a classifier to distinguish between indoor and outdoor images.

A. Low-Level Features

Tab. 1 lists the low-level features used for image indoor/outdoor classification. Color information provide important cues for indoor/outdoor classification, like the outdoor images usually contain sky part which means color blue take up the large part in the color distribution. For the first color feature C1, a global color histogram is computed by dividing RGB color space into 27 bins with each component in three ranges. Then, we transform the image into the YCbCr color space, divide it into seven horizontal bands, and compute the mean (C2) and the standard deviation (C3) of each of the three color.

Besides color information, we also extract edge information. To describe the edges distribution in the scene, we used an 18 bin edge direction histogram (E1). The gradient of the luminance image is computed using Gaussian derivative filters, and the gradient with the magnitude exceeding a set threshold contribute to the histogram.

Texture information is another important cues. We extract texture information based on a multi-resolution analysis and get ten different sub-bands by computing the three level wavelet transform of the luminance image. For each band, we then compute the average absolute value of the coefficients (T1) and their standard deviation (T2). Finally, each image is represented by a feature vector of $27+21+21+18+10+10 = 107$ dimension.

B. Classifier

After finishing extracting features from images, the features are fed into some classifier to distinguish indoor and outdoor images. There exists a lot of classifiers like SVM, random forests. In Schettini et la. [3], several classification strategies were designed and experimentally compared, and they found that random forests can provide reasonably good performance and robustness. In our pipeline, we also use random forest to classify the images into indoor and outdoor categories.

Random forests are an ensemble learning method for classification and they construct a multitude of decision trees at training time. Decision trees are produced by recursively partitioning a training set of feature vectors labeled with the correct class. Each split consists of a comparison between the value of a single feature and a threshold. A class is assigned to each of the terminal nodes. During testing time, the value of a single feature is compared with the fix threshold at each split for each decision tree, and its predicted class is the class associated with the terminal node it reaches. All the decision trees from the random forests will vote for the final result. While using random forests to classification, no a prior knowledge about the distribution of the features is needed and the issue of feature normalization can be ignored. Similar with [3], we used decision trees built according to the CART methodology [5] and set the size of decision trees to 50 trees.

TABLE I
LOW-LEVEL FEATURES FOR INDOOR/OUTDOOR CLASSIFICATION.

Feature Descriptions	Num
C1. RGB values: histogram (27 bin)	27
C2. YCbCr values: mean (7 bands)	21
C3. YCbCr values: standard deviation (7 bands)	21
E1. Edge direction values: histogram (18 bin)	18
T1. coeff. of wavelet transform: mean (10 bands)	10
T2. coeff. of wavelet transform: standard deviation (10 bands)	10

III. Mid-Level Cues from Superpixels

In order to extract mid-level cues, superpixels are used as process units, which will add a natural spatial support to compute features. There have been a lot of the methods to segment an image in a bottom-up way such as Normalized Cuts (NCuts) [6], the Felzenszwalb and Huttenlocher (FH) [7] and Mean-Shift [8]. Felzenszwalb and Huttenlocher is typically used in high recall settings to create a gross oversegmentation into thousands of superpixels, while NCuts provides better precisions.

In this paper, we want to obtain multiple segmentations of an image into geometrically homogeneous regions. Given an image, we first generate *multiple segmentations* per image. We use Normalized Cuts for segmentation and vary the number of segments from 3 to 12 to obtain 75 segments per image. Fig. 1 shows three examples of multiple segmentations. We can find that most of the superpixels are objects or contain parts of the objects. Thus, we can take superpixels as process units, and build features based on the distribution of superpixels. To model the distribution of superpixels, we use the bag-of-word method to encode the superpixels from each image and generate a fix length feature vector using histogram.

To build the visual words, from all the superpixels, we extract several types of bag-of-features histograms, such as Texton Histograms, Color Histograms, and pyramid of HOG.

For Texton Histograms, we use a filter bank with 18 bar and edge filters (6 orientations and 3 scales for each), 1 Gaussian, and 1 Laplacian-of-Gaussian filters. We quantize to 400 textons via k-means. For Color Histograms, we use Lab color space, with 23 bins per channel. For pyramid of HOG, we use 3 pyramid levels with 8 bins. Each features are normalized to sum to 1. We randomly select 10000 segments and cluster these segments to form 190 visual words. For each superpixel of the image, it will be assigned a visual word that is most close to it. And the image is represented by the histogram of the visual words.

IV. Experiments

In this section, we analyze and evaluate the proposed method on a real-world data set. We will describe the dataset used for evaluation in subsection IV-A, then show the results of the classification on this dataset in subsection IV-B.

A. Dataset

The data set collected by Gehler et al. [9] contains 246 indoor images and 322 outdoor images. We evaluate our indoor/outdoor classification method on this data set, using three-fold cross-validation. The random forest models are learned using 2 of 3 folds, and the algorithm is tested on the remaining fold. The whole procedure is repeated 3 times, leaving out each fold once for testing.

B. Classification Results

In this section, four experiments are conducted on the data set collected by Gehler et al. [9]. In the first experiment, we train the random forests model using only low-level features. Then, we obtain the following results for indoor/outdoor classification (see Tab.2).

TABLE II
Confusion matrix obtained on the dataset collected by Gehler et al. [9] using only low-level features.

	Predicted indoor	Predicted outdoor
True indoor	83.4%	16.6%
True outdoor	9.6%	90.4%

The overall classification accuracy obtained on this data set is 87.3%. The number of misclassifications is 72 (41 indoor and 31 outdoor images).

In the second experiment, the random forests model is trained using only mid-level features, and the results for indoor/outdoor classification is shown in Tab. 3. The overall classification accuracy is 83.8% and the number of misclassifications is 92 (67 indoor and 25 outdoor images). When compared with the method using only low-level features, the classifying rate for outdoor images get improved. The main reason may be because the visual words we build for mid-level features are more discriminative for outdoor images.

In the third experiment, we train the random forests model using both low- and mid-level features. The feature vector of low-level features and the vector of mid-level features are concatenated together to a new feature vector. Then, we obtain the following results for indoor/outdoor classification (see Tab.4).

TABLE III
Confusion matrix obtained on the dataset collected by Gehler et al. [9] using only mid-level features.

	Predicted indoor	Predicted outdoor
True indoor	72.6%	27.4%
True outdoor	7.9%	92.1%

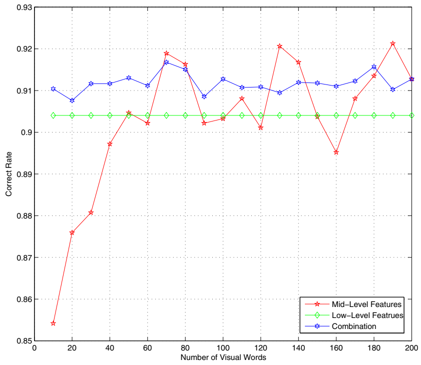

Fig. 4. Correct rate for outdoor images varies with the number of visual words changing.

TABLE IV
Confusion matrix obtained on the dataset collected by Gehler et al. [9] using low- and mid-level features.

	Predicted indoor	Predicted outdoor
True indoor	84.2%	15.8%
True outdoor	8.4%	91.6%

The overall classification accuracy is 88.4% and the number of misclassifications is 66 (39 indoor and 27 outdoor images). It can be found that the classifying rate for outdoor images and indoor images are both been improved. Fig. 3 shows samples of misclassified images. These images are very difficult to classify,even for human. They usually have very similar information about the context in which they are taken with the categories that they are misclassified.

In the forth experiment, we will analysis whether the classification rate is sensitive to the number of visual words. We vary the number of the visual words K from 10 to 200 with 10 each step. Fig. 4 shows how the correct rate for outdoor images varies with the K value changing, and Fig. 5 shows how the correct rate varies for indoor images.

Overall, from the Fig. 4, we can find that the correct rates for outdoor images grow with the number of visual words increasing . The rates start to reach the correct rate

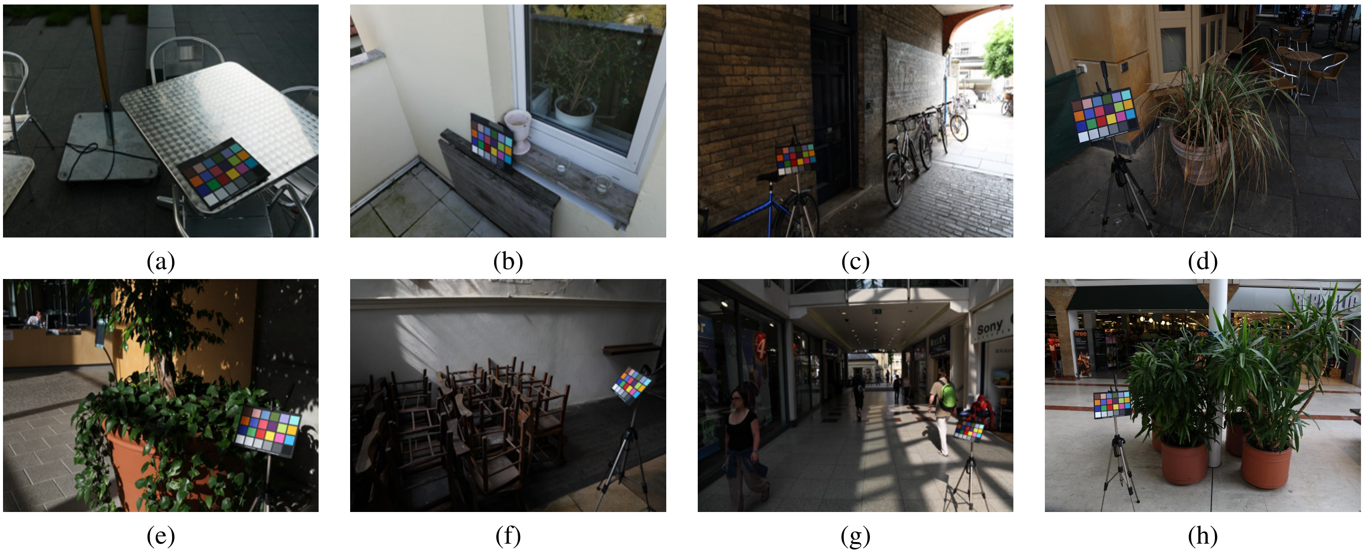

Fig. 3. Example of outdoor images misclassified as indoor (a-d) and some indoor images misclassified as outdoor images (e-h)

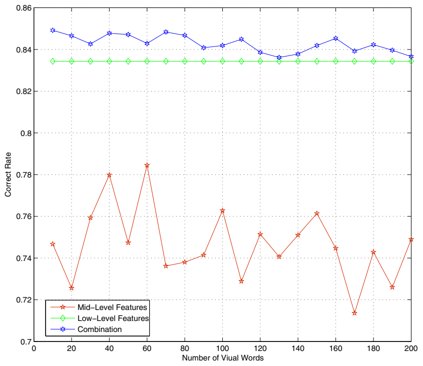

Fig. 5. Correct rate for indoor images varies with the number of visual words changing.

using only low-level features with number of visual words equalling to 50. When the number of visual words is above 50, there exist oscillations along the correct rates using only low-level features. However, the correct rates using both low-level features and mid-level features always perform better than that using only low-level features.

For indoor images, the correct rates using mid-level features seems not like to vary with the changing of the number of visual words, and always perform worse than that using only low-level features. However, the correct rates using both low-level features and mid-level features always perform better than that using only low-level features. From Fig. 4 and Fig. 5, it is can be found that the correct rates for outdoor images and indoor images both got improved, which verifies our hypothesize that pixel based low-level descriptions are useful but can be improved with the introduction of mid-level region information.

V. Conclusions

In this paper, we focus on the indoor/outdoor images classification with the emphasis on the feature extraction step. We hypothesize that pixel based low-level descriptions are useful but can be improved with the introduction of mid-level region information. The feature description is explored by using spatial information from mid-level cues to obtain a superior scene description. Therefore, we introduce a superpixel-based region descriptor and apply it to indoor/outdoor image classification. Based on the experimental evaluations, we can verify our hypothesis that mid-level cues enrich the image description and improve the performance of low-level cues. For further research, we will try to apply mid-level features into other image category classification work, like stage classification.

Acknowledgment

We would like to thank the reviewers for their insightful and constructive comments. This work was supported by Shandong Provincial Natural Science Foundation of China (Grant No. ZR2011FM037) and Innovation Fund for Distinguished Graduate Student of Shandong University (Grant No. yyx10043).

References

[1] S. Bianco, G. Ciocca, C. Cusano and R. Schettini, “Improving Color Constancy Using Indoor-Outdoor Image Classification,” *IEEE Transactions on Image Processing,* pp. 2381-2392, 2008.

TABLE V
Summary of the features used to describe the patches.

Feature Descriptions	**Num**
C1. RGB color histogram	27
C2. YCbCr color moments	6
C3. Cast indexes	2
C4. Number of colors	1
E1. Edge magnitude histogram	5
E2. Edge direction histogram	18
T1. Wavelet statistics	20

[2] M. Szummer and R. Picard, “Indoor-Outdoor Image Classification,” *in Proc. Int. Workshop Content-Based access of Image and Video Databases,* pp. 42-51, 1998.

[3] R. Schettini, C. Brambilla, C. Cusano and G. Ciocca, “Automatic Classification of Digital Photographs Based on Decision Forests,” *Int. J. Pattern Recognit. Artif. Intell.,* vol. 10, no. 1, pp. 819-845, 2004.

[4] A. Oliva and A. Torralba, “Modeling the shapes of the scene: a holistic representation of the spatial envelope,” *IJCV,* vol. 42, no. 3, pp. 145-175, 2001.

[5] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, “Classification and Regresion Trees,” *New York: Wadsworth and Brooks/Cole,* 1984.

[6] S. Jianbo and M. Jitendra, “Normalized Cuts and Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 22, no. 8, pp. 888-905, 2000.

[7] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient Graph-Based Image Segmentation,” *Int. J. Comput. Vision,* vol. 59, no. 2, pp. 167-181, 2004.

[8] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, 2002.

[9] P. V. Gehler, C. Rother, A. Blake, T. Minka and T. Sharp, “Bayesian Color Constancy Revisited,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008),* pp. 1-8, 2008.